answer open-ended questions included in the 1999 study were the meaning of various scientific terms such as DNA, molecule, and radiation. Short answer open-ended questions were coded on an on-going basis, while long answer open-ended questions were coded in three batches, after 1,000 completions, 1,500 completions, and 1,882 completions.

Each week Dr. Kimmel prepared files containing the responses for all short answer open-ended questions collected during the previous week. The Electronic Coder (EC) software program was used by Research Assistants to code all short answer open-ended questions. Short answer questions were coded in a relatively mechanized manner by identifying key words in the respondents' answers. After the first week of coding, the Research Assistants provided Dr. Kimmel with hard copies of the codes and actual responses for each open-ended question. Dr. Kimmel reviewed the codes with the Research Assistants and made changes where needed in the codes assigned, and provided additional training as needed in the use of the EC and the codes to be assigned. The codes assigned to the short-answer questions were merged with the main SPSS analysis file on a weekly basis. At the conclusion of the study Dr. Kimmel reviewed all of the codes assigned for each short answer open-ended question.

Questions classified as long answer open-ended questions varied in length between 74 and 560 characters. These responses cannot be coded using the EC, for they require complex judgments by the coders. Three graduate students in biology and chemistry were hired to code the long answer open-ended questions included in the 1999 study. All of the graduate students had also coded some of the same open-ended responses for two previous surveys conducted by the International Center for the Advancement of Scientific Literacy. Dr. Kimmel reviewed the coding categories with the coders at the beginning of the survey, and provided the coders with written coding instructions, and examples containing responses assigned each possible code from past surveys for each of the questions. The coders worked independently, and were not allowed to discuss their coding assignments with the other coders. At the conclusion of the survey, at least Dr. Kimmel reviewed all inter-coder disputes (cases in which the three coders were not in agreement) and assigned a final code to those cases.

Construction of Scales and Other Summary Measures

A number of constructed variables were created at the conclusion of data collection for use in the preparation of the appendix tables for ***Science and Engineering Indicators.*** The major constructed variables include attentiveness to selected public policy issues, exposure to formal science and mathematics education, and various categorizations of the respondents' occupation, education, and age. (See Appendix I for the SPSS commands used to create all constructed variables.)

QUALITY PROFILE

Review of Questions

Several pieces of information were included with each case to assist with the review of data quality. First, interviewers were asked to rate each respondent's comprehension of the questions included in the study, by assigning a code of high, moderate, or low to each case. A total of 113 cases, or 6 percent of the sample, were assigned a low comprehension rating by the interviewers. Dr. Kimmel reviewed each

of these 113 cases, to determine if it appeared that any of these respondents had substantial problems understanding the questions. After a careful review of these cases, none were excluded from the final analysis file. In general, the interviewers had assigned a code of low comprehension if the respondents answered many of the knowledge-based questions incorrectly. However, none of the 113 cases exhibited unusual response patterns, and as such were included in the fmal analysis file.

Second, interviewers were asked to how seriously each respondent appeared to take the interview, by assigning a code of very serious, moderately serious, or not serious. Only six respondents were rated as not being serious in completing the interview. Dr. Kimmel produced a separate codebook for each of these six respondents, and examined the response patterns for potential problems. On the basis of this review, none of these six cases were excluded from the final analysis file, as none of these cases exhibited unusual response patterns.

Item Non-response for Critical Items

A total of 1,884 interviews were completed for the study. Five questions have historically been treated as critical items for inclusion in *the* final analysis files for *the Science and Engineering Indicators* studies, and thus have complete response. The five items-number of adults in the household, gender, age, level of education, and race-ethnicity-are all required to create the analytic weight that is used for analyses. Two of the completed interviews were dropped from the analysis file because they were missing one or more of the critical items required to weight the data file (number of adults in the household, age, level of education, or race-ethnicity). These two records have been dropped from all analysis files and are not included in the final files that have been forwarded to the National Science Foundation. The remaining 1,882 completed interviews (99.9 percent of the total) contained complete responses on all of the critical items.

All other questions in the study are treated as non-critical items. There was only minimal non-response for non-critical items in the survey. It should be noted that a "don't know" response is not considered to be non-response. Only actual refusals to respond are considered to be non-response. The item with the highest level of non-response, was a question in which the respondents were asked to agree or disagree with the statement, "We depend too much on science and not enough on faith." Even for this item, only .8 percent of the sample refused to give a response. No values were imputed for any items with non-response, but are coded as "won't say" in the data file. Table 10 provides a summary of all non-critical items with some non-response.

Table 10. Distribution of Item Non-response.

Question Number	Question	Number of Cases	Unweighted Percent
9	INTEREST - AGRICULTURAL & FARM	1	0.1
15	INTEREST - NEW MEDICAL DISCOVERIES	1	0.1
17	INTEREST - ENVIRONMENTAL POLLUTION	1	0.1
23	INFORMED - ECONOMIC & BUSINESS CONDITIONS	1	0.1
54	NUMBER OF VISITS ART MUSEUM	1	0.1
55	NUMBER OF VISITS NATURAL HISTORY MUSEUM	1	0.1
56	NUMBER OF VISITS ZOO OR AQUARIUM	1	0.1
57	N UMBER OF VISITS SCI/TECH M US EU M	1	0.1
58b	BORROW BOOKS LIBRARY YES/NO	1	0.1
59a	DID R BUY ANY BOOKS	1	0.1
60	WORLD BETTER OR WORSE DUE TO SCI	4	0.2
65	UNDERSTANDING OF 'DNA'	1	0.1
79	EXPERIMENTAL METHOD	8	0.4
83	SCI & MATH ED IN U.S. IS INADEQUATE	6	0.3
85	DEPEND TOO MUCH ON SCI-NOT ENUF ON FAITH	15	0.8
86	FED GOVT SUPPORT ALL SCI RSRCH	6	0.3
88	SCI NOT IMPORTANT IN EVERYDAY LIFE	1	0.1
89	BUILD SPACE STATION TO HOUSE EXPERIMENTS	7	0.4
90	SOME NUMBERS LUCKY FOR SOME PEOPLE	6	0.3
91	SCI MAKES LIFE CHANGE TOO FAST	1	0.1
92	SCI WANTS TO MAKE LIFE BETTER	2	0.1
93	TECH DISCOVERIES WILL DESTROY THE EARTH	2	0.1
95	MORE OPPORTUNITY FOR NEXT GENERATION	1	0.1
96	TECH CREATES INHUMAN WAY OF LIFE	1	0.1
97a	NEW INVENTNS ALWAYS FOUND FOR BAD THINGS	4	0.2
97b	PEOPLE DO BETTER LIVING SIMPLER LIFE	3	0.2
98	BENEFIT-HARM BALANCE SCIENTIFIC RESEARCH	2	0.1
99	DEGREE OF BENEFIT BALANCE-SCI RESEARCH	1	0.1
101	RISK BENEFIT GENETIC RESEARCH--BOTH	12	0.6
103	DEGREE OF RISK BALANCE-GENETIC RESEARCH	1	0.1
104	RISK & BENEFIT IN NUCLEAR REACTORS	3	0.2
105	DEGREE OF BENEFIT BALANCE-NCLR REACTORS	1	0.1
107	COST BENEFIT BALANCE-SPACE PROGRAM	1	0.1
111	GOVT SPENDING ON REDUCING POLLUTION	1	0.1
112	GOVT SPENDING ON IMPROVING HEALTH CARE	2	0.1

Table 10. Distribution of Item Non-response: continued

Question Number	Question	Number of Cases	Unweighted Percent
113	GOVT SPENDING ON SUPPORTING SCI RESEARCH	2	0.1
114	GOVT SPENDING ON IMPROVING EDUCATION	3	0.2
115	GOVT SPENDING ON HELPING OLDER PEOPLE	2	0.1
116	GOVT SPENDING ON IMPROVING NTNL DEFENSE	1	0.1
117	GOVT SPENDING ON HELPING LOW INCOME	2	0.1
121	RADIOACTIVITY MAN-MADE	2	0.1
124	LASERS FOCUS SOUND WAVES	2	0.1
125	ELECTRONS SMALLER THAN ATOMS	1	0.1
127	UNIVERSE BEGAN WITH EXPLOSION	6	0.3
128	CONTINENTS MOVING & WILL CONTINUE TO DO	4	0.2
129	HUMANS DEVELOPED FROM EARLIER SPECIES	3	0.2
130	SMOKING CAUSES LUNG CANCER	1	0.1
131	HUMANS AND DINOSAURS COEXISTED	1	0.1
132	RADIOACTIVE MILK MADE SAFE BY BOILING	2	0.1
135	LENGTH OF EARTH ORBIT	1	0.1
137	PROBABILITY - LAST THREE WILL NOT	1	0.1
138	PROBABILITY - SAME RISK FOR EACH	1	0.1
139	PROBABILITY -NONE IF ONLY THREE	2	0.1
142	REGARD FOR ASTROLOGY REPORT	2	0.1
147	MARITAL STATUS	7	0.4
148	/NUMBER OF CHILDREN	3	0.2
149	NUMBER OF CHILDREN AT HOME	5	0.3
151	FIELD OF HIGHEST DEGREE ATTAINED	2	0.1
152	NUMBER OF COLLEGE SCIENCE COURSES	1	0.1
153	HIGHEST LEVEL MATH IN HS	2	0.1
155	TOOK HS BIOLOGY COURSE	1	0.1
156	TOOK HS CHEMISTRY COURSE	1	0.1
157	TOOK HS PHYSICS COURSE	1	0.1
160	EMPLOYMENT STATUS	2	0.1
161	RESPONDENT'S OCCUPATION CODE (CENSUS 70)	9	0.5
166b	MORE THAN ONE COMPUTER IN HH	1	0.1
170	R HAS CD-ROM READER AT HOME	1	0.1
171	R HAS MODEM AT HOME	1	0.1
174d	HAVE WEB TV AT HOME	1	0.1
175c	GET INFO FROM WWW OR BROWSE	2	0.1
176a	R SMOKES	1	0.1

Table 10: Distribution of Item Non-response: Continued

Question Number	Question	Number of cases	Unweigbted Percent
177b	COMMUNITY TYPE	1	0.1
181	RESPONDENT YEAR OF BIRTH	10	**0.5**
dog	LAB DOG/CHIMP OK IF NEW INFO FOUND	13	**0.7**

Results of Monitoring of Interviewers

An important mechanism of quality control for all NORC telephone studies is the monitoring system. The computerized monitoring system randomly selects interviewers who are on-line and feeds the call to a supervisor who can silently monitor the call from their station. Monitoring takes place whenever the phone study is active. The system also allows the CATI data capture to be followed in real time on the supervisor's screen. That is, the supervisor can hear the telephone conversation and see what the interviewer is recording. Another feature allows off-site authorized parties, such as the Chicago Academy of Sciences staff or NSF staff, to monitor the calls.

All interviewers on the 1999 study were monitored heavily at the inception of the project. Supervisors made notes of areas where interviewers departed from standardized procedures or had other problems. The supervisors met with project managers often in the early days of the study to discuss problem areas across the interviewers and to establish correctional procedures. Questions asked by respondents, that had not been anticipated and included in the training materials, were reviewed and scripts were developed.

The supervisors met individually with each interviewer to provide feedback and to coach them on technique. Overall, most interviewers did very well with the technical terminology of the instrument and on the precision needed to capture the verbatim remarks correctly. All the interviewers who were trained for the 1999 study passed the certification period, which included the results of their monitoring sessions. In addition to the ongoing monitoring, the Chicago Academy of Sciences staff reviewed the early data returns to see that the interviewers were correctly capturing the survey data.

As cases became more difficult in the later stages of the data collection, monitoring was used as a way to examine how certain interviewers were particularly successful in converting respondents. Their techniques were then shared with the other interviewers in group debriefings and brainstorming sessions.

A weekly monitoring report was produced which showed, for each interviewer, the number of hours worked and number of hours and percent of time they were monitored. The report also showed the cumulative time for each category. The NORC quality control system minimally monitors 10 percent of each interviewer's completed cases.

Because the verbatim responses were particularly important in this study, the NORC project managers and the Chicago Academy of Sciences staff often monitored cases to assure that high quality was maintained. Any Chicago Academy of Sciences feedback to NORC from the monitoring sessions was shared with the supervisors and disseminated via meetings or revised protocols to the interviewers.

Review of Coding Error Rates

Coders were instructed to assign the full range of codes for each of the long open-ended questions. The number of possible categories to be assigned ranged from a low of four for the meaning of the terms *Internet* and *molecule* to a high of seven for *scientific study*. In broad terms, the coders tended to have higher levels of agreement when there were fewer coding categories to assign. Although there were

actually eight possible codes that could be assigned for the experiment question[7] the categories are linked to the respondent's choice of one or two groups in the preceding question. As a result of this linkage, the coders actually were able to assign only four possible choices for any given response to the experiment question. The full categories are generally collapsed in analyses into two categories — a correct understanding of the concept or an incorrect understanding.

When the inter-coder agreements are examined for the full range of categories, the level of agreement ranged from a low of .34 (for coders one and three) for the meaning of scientific study to a high of .81 (for coders one and two) for the experiment question (see Table 8)[8]. The three coders were in complete agreement on the code to assign to the experiment question for 76 percent of the cases, and for 75 percent of the cases for the Internet question. The three coders were in agreement about the code to be assigned for 50 percent of the DNA responses, and two of the coders were in agreement for another 44 percent of the responses. In 57 percent of the cases, all three of the coders assigned the same code to the molecule responses, and in another 37 percent of the cases two of the coders assigned the same code. The three coders assigned the same code for 46 percent of the responses to the radiation question, and in another 42 percent of the cases two of the coders were in agreement. In 35 percent of the cases the coders were in complete agreement regarding the scientific study question, and in another 52 percent of the cases, two of the coders were in agreement.

Few of the inter-coder disagreements involved conflicts over whether a response was scientifically correct or incorrect. Instead, the majority of the disagreements revolved around which incorrect response was most appropriate. When the coding categories are collapsed into scientifically correct or incorrect responses, the inter-coder agreement measures increase substantially. Half of the possible kappas are substantial, ranging between .61 and .80, and another 28 percent are almost perfect, exceeding the value of .81. The three coders were in complete agreement for 90 percent of the cases for the molecule question, 88 percent for the DNA, 85 percent for the Internet and radiation questions, 82 percent for the experiment question, and 78 percent of the cases for the scientific study question (see Table 11).

All of the coding was closely supervised, as has been described in previous sections, and the inter-coder agreements, as measured both by kappa and percent of agreement, were high. Nonetheless, it is likely that some of the responses were not coded accurately by individual coders, accounting for some of the lower kappas. For exactly this reason, all inter-coder disagreements (cases in which the three coders were not in agreement) were reviewed by Dr. Kimmel, who assigned a final code to those cases based on the coding protocols that were established for the study by Professor Miller, and that have been used in previous *science and engineering indicators* studies. Thus, although the kappas *were* high, in particular with regards to the dichotomous measures of knowledge (scientifically correct or incorrect), the final codes that were assigned, and that are present in the data set, reflect an even higher level of data quality.

---

[7] Now, please think about this situation. Two scientists want to know if a certain drug is effective against high blood pressure. The first scientist wants to give the drug to 1000 people with high blood pressure and see how many of them experience lower blood pressure levels. The second scientist wants to give the drug to 500 people with high blood pressure, and not give the drug to another 500 people with high blood pressure, and see how many in both groups experience lower blood pressure levels. Which is the better way to test this drug? Why is it better to test the drug this way?

[8] Kappas between .O to .20 are considered to be slight, between .21 and .40 to be fair, between .41 and .60 to be moderate, between .6 1 and .80 to be substantial, and .8 1 and higher to be almost perfect. For more information about kappa as a measure of observer agreement, see Landis & Koch (1977) The measurement of observer agreement for categorical data, *Biometrics, 33,* p. 159, or Fleiss (1981) *Statistical Methods for Rates and Proportions* (2nd Edition), New York: Wiley.

Table 11. Inter-coder agreement, 1999 study.

Question	Categories	Kappa between coders 1, 2, and 3			Percent of Inter-coder agreement		
		1 & 2	1 & 3	2 & 3	3 agree	2 agree	0 agree
DNA	all categories	.60	.52	.53	50%	44%	6%
DNA	correct/incorrect	.86	.82	.84	88	12	--
Scientific study	all categories	.52	.34	.42	35	52	13
Scientific study	correct/incorrect	.70	.58	.68	78	22	--
Internet	all categories	.70	.68	.70	75	23	1
Internet	correct/incorrect	.57	.58	.58	85	15	--
Molecule	all categories	.62	.56	.54	57	37	6
Molecule	correct/incorrect	.82	.82	.79	90	10	--
Radiation	all categories	.59	.50	.52	46	42	12
Radiation	correct/incorrect	.73	.66	.66	85	15	--
Experiment	all categories	.81	.77	.78	76	23	1
Experiment	correct/incorrect	.78	.73	.74	82	18	--

LIMITATIONS

In all survey research there are several possible sources of error, and it is important to recognize these possible sources of error. The primary sources of error connected to the 1999 study are discussed in the following paragraphs.

One source of error in the 1999 study is the exclusion of households without telephones. Approximately 95 percent of households in the United States (excluding group quarters like dormitories, prisons, and hospitals) have a telephone. The presence of a telephone is lowest among low-income individuals and families and among non-English speaking groups. Even though the weight procedure makes some correction for this kind of error, it is likely that even within weighting cells, individuals with and without phones may have slightly different attitudinal or demographic profiles.

A 'second source of error in the 1999 study is the refusal to participate. As was discussed above, approximately a third of possible respondents either directly refused to participate or were able to use an answer machine or other device to avoid talking to an interviewer. Some of this distortion may be corrected by weighting, but it is likely that some of the differences between those who are willing to talk and those who are not is not corrected by weighting.

item comprehension is a third potential source of error in the study. It is likely that some respondents, especially those individuals with little formal education or little exposure to science, may not have fully